TERDOO ACHU

CSCE 581

Quiz 1 Report (COMPAS Fairness Analysis)

1. Objective

The goal of this project was to analyze the fairness and performance of machine learning classifiers using the COMPAS recidivism dataset. The task involved training classification models to predict whether an individual would reoffend within two years and evaluating both accuracy and potential bias across demographic groups.

The project emphasizes Trusted AI principles by examining model errors and how performance varies across race categories.

2. Dataset Description

The COMPAS dataset contains demographic and criminal history information for individuals assessed by a risk prediction system. The target variable used in this project was:

two_year_recid

where:

- **0** = no recidivism
- **1** = recidivism within two years

Relevant features included:

- age
- sex
- race
- prior counts
- risk scores

The race categories used were:

- African-American
- Asian
- Caucasian

- Hispanic
- Native American
- Other

3. Data Processing

Before training models, the dataset was cleaned and prepared:

- Selected relevant columns for prediction
- Encoded categorical variables using label encoding
- Split data into training and testing sets (80% train / 20% test)
- Standard preprocessing applied to ensure compatibility with scikit-learn models

4. Classification Models Used

Two classification algorithms were trained:

4.1 Logistic Regression

Logistic Regression was used as a baseline linear classifier. It provides a simple and interpretable model for binary prediction tasks.

4.2 Decision Tree Classifier

The Decision Tree model was used as a second classifier. Decision Trees were chosen because they are interpretable and allow analysis of decision patterns, which is important for fairness analysis.

5. Model Performance

Logistic Regression Results

Accuracy: **69%**

- Precision (class 0): 0.71
- Recall (class 0): 0.76
- Precision (class 1): 0.65
- Recall (class 1): 0.59

The model showed stronger performance in predicting non-recidivism compared to recidivism.

Decision Tree Results

Accuracy: **69%**

- Precision (class 0): 0.72
- Recall (class 0): 0.75
- Precision (class 1): 0.65
- Recall (class 1): 0.61

Performance between models was similar, although the Decision Tree slightly improved recall for recidivism cases.

6. Confusion Matrix Analysis

The confusion matrix revealed:

- False positives: individuals predicted to reoffend who did not
- False negatives: individuals predicted safe who reoffended

These errors are important because they can lead to unfair outcomes in real-world decision systems.

7. Fairness and Bias Analysis

Error rates were computed across race groups to evaluate fairness.

Example findings:

- Native American group showed the highest error rate (~ 66%)
- Asian group ~ 40%
- African-American group ~ 31%

This indicates that the model does not perform equally across demographic groups. Unequal error rates suggest potential bias or imbalance in the data, highlighting the importance of fairness evaluation in AI systems.

8. Discussion

Although both classifiers achieved similar accuracy (~ 69%), fairness analysis revealed differences in model behavior across race categories. This demonstrates that high accuracy alone does not guarantee equitable outcomes.

Trusted AI requires evaluating:

- Model accuracy
- Error distribution
- Demographic fairness

The COMPAS dataset is a well-known example where machine learning predictions can raise ethical concerns.

9. Conclusion

This project demonstrated how machine learning models can be evaluated not only for performance but also for fairness. Logistic Regression and Decision Tree classifiers achieved similar results, but analysis of group error rates revealed disparities across races.

Future work could include:

- Bias mitigation techniques
- Feature balancing
- Alternative fairness metrics

This experiment shows that machine learning systems must be evaluated beyond accuracy metrics, especially when used in socially sensitive domains such as criminal justice.